Climbing Mount Dependable: How to detect misinformation on the Internet

Nicolaisen, Jeppe

Published in:
Proceedings of the NETs2013 International Conference on Internet Studies

Publication date:
2013

Document version
Early version, also known as pre-print

International Conference on Internet Studies,
September 7-8, 2013, Hong Kong, China

CLIMBING MOUNT DEPENDABLE: HOW TO DETECT MISINFORMATION ON THE INTERNET

Associate Professor Jeppe Nicolaisen,
Royal School of Library and Information Science, Denmark
jni@iva.dk

ABSTRACT

The open nature of the Internet makes it vulnerable to misinformation. When seeking and browsing the Internet it is therefore important to check the validity of retrieved information. Various techniques have been proposed for critically assessing Internet sources and information. The paper presents and reviews two of these: *the checklist method* and *comparative evaluation*. The checklist method is found to be too inconsistent to adequately function as a dependable technique. Comparative evaluation is found to be a more dependable technique if supplemented with knowledge about document typology and the so-called authority hierarchy. Using a real example from actual information seeking on the Internet, the paper shows how to detect misinformation using knowledge about document typology and comparative evaluation.

We cannot direct the wind,
but we can adjust the sails.
Bertha Calloway

1. INTRODUCTION

The Internet is truly fantastic. At the World Wide Web we can get in contact with people from all over the world and with fellow people who share our interests. We can search and find information about all kinds of topics. We can learn about new things, more about things we already know something about, and thus become more informed and knowledgeable about the world we live in. Moreover, we can also participate ourselves with our own knowledge and ideas. We can produce homepages, join the ongoing debate on discussion forums and on social networks, create our own blogs, edit pages on Wikipedia, etcetera, etcetera. In short: The Internet is truly fantastic!

The Internet is distinctively a democratic medium. If you have a device with Internet access you can in principle join and participate. This openness is a huge plus. That all in principle can join and participate in the ongoing discussion makes it possible for many views and arguments to be heard. However, this openness comes at a price: Misinformation. There are, unfortunately, many examples of misinformation on the Internet having caused more or less fatal consequences. Among the lighter and fun ones is the politician from New Zealand who demanded a total ban on dihydrogen monoxide [1]. On the Internet she had learned that this

substance is extremely poisonous and dangerous. Dihydrogen monoxide is, however, the chemical name for plain and simple water (H_2O). Another amusing example is the sports journalist from a British newspaper who traveled to Cyprus to cover a soccer match between a Cypriote and an English team. In the article he wrote to his newspaper he reported among other things that the Cypriote fans were known as the 'Zany Ones' and that they wore hats made from shoes. Later, it was proven that the journalist in question had gotten the information not from meeting the fans himself, but from a page on Wikipedia that had been edited by rivaling fans [2].

There are, however, also examples of misinformation having caused severe injury and even death [3]. Medical misinformation on the Internet has thus led to severe consequences for a number of people having followed advice on self-medication found on various Internet sites.

As an information seeker and user of the Internet it is consequently important or even vital to keep in mind that not all information searched and found is accurate and true. One should of course always meet new ideas and information with a sound portion of skepticism – not just on the Internet [4, 5]. However, a vital difference between the Internet and traditional media like books, newspapers, journals, and magazines, is that peer review and quality control is often absent on the Internet. Information on the Internet has in other words only very infrequently gone through the same instances of quality control as e.g. a book published by an esteemed publisher or like papers published by academic presses. One should consequently stay even more alert to the information picked up on the Internet. The aim of this paper is to present and discuss various techniques for separating the weed from the chaff when browsing and seeking the Internet. Such techniques are sometimes referred to as techniques for *source criticism*. However, in order to become a true expert in source criticism, this paper will argue that additional knowledge and expertise in document typology is also needed. Different types of documents play different kinds of roles in the global communication system – or perhaps more precise, they conduct different functions with varying degrees of authority. Knowledge about these elements is essential in order to be able to navigate safely on the Internet. A section on document typology and authority consequently follows the section on techniques for source criticism. The paper ends with an example of how to use techniques for source criticism and knowledge about document typology to assess the accuracy of a knowledge claim found on the Internet: The so-called *open-access advantage postulate*.

2. TECHNIQUES

Seek not greatness, but seek truth
and you will find both.
Horace Mann

There are various techniques for source criticism available to the serious information seeker. We will start by examining the so-called *checklist method* and later move on to examine the so-called *comparative evaluation method*.

2.1 Checklists

In short, the checklist method simply instructs the searcher to evaluate the source in question by comparing the source to a list of questions formulated as a checklist. The answers to these questions are thus believed to indicate to what extent the searcher may trust the information

provided by the source in question. There exist many checklists [6]. They are typically provided by libraries and information centers in an attempt to help their users to navigate the Internet more safely. Although there is no agreed upon standard, most checklists seem to be addressing more or less the same categories of questions. Typically these categories include:

- Questions on authority
- Questions on accuracy
- Questions on objectivity
- Questions on currency
- Questions on coverage

Within each category a number of questions are asked. Zooming in on the questions on authority, these are the typical questions:

- Who is the author?
- What are his or her qualifications?
- Is the author affiliated with an institution?
- Are there contact information?
- Are there advertisements and/or sponsor influence?

The rationale behind these questions is that if an author is willing to put his or her name on the source, and if it is even possible to detect that s/he has qualifications in the area that make him or her a trustworthy authority on these matters – maybe even is affiliated with a honorable institution – then we may trust the source and its information to a larger extent than if not (especially if the page is free from advertisements and sponsor influence).

Undoubtedly, the checklist method can catch some sources that consciously or unconsciously misinform their readers. Yet, the method is not error free. It lets too many sources through that should not be trusted, and conversely stops too many sources that actually are quite reliable. One example of such a source that slips through the eye of the needle of most checklists is the homepage entitled *'Martin Luther King Jr. – A True Historical examination'.* It is located at www.martinlutherking.org, and contains a number of texts about Martin Luther King. One of the texts is a resource guide for students writing reports about King, and one of the links in the guide is a link to a two-page article entitled *'The Beast as Saint'.* The main points of the article may be summarized quoting a few lines from the article:

> "Well friends, he is not a legitimate reverend, he is not a bona fide PhD, and his name isn't really "Martin Luther King, Jr." What's left? Just a sexual degenerate, an America-hating Communist, and a criminal betrayer of even the interests of his own people" [7].

The article ends with a lists of references used in the text and contact information (the author and his homepage). Despite *'Martin Luther King Jr. – A True Historical examination'* scores high on most checklists (it is located on a dependable .org-address, we can find contact information, we can find references to cited works in the text, etc.), most people would definitely assess the source to be straight out untrustworthy after just little more detailed inspection.

Another problem with most checklists was pointed out first by Marc Meola [8]. Often, when dealing with the questions on accuracy, the checklists ask:

- Is the information reliable and error-free?

As noted by Meola, this question is at best a little naïve. To answer this very question is precisely why the checklist method was applied in the first place. If one could answer the question straight up, one would not need a checklist or any other method. Consequently, it makes no sense to ask the question without further instructions on how the information seeker is to go about answering it.

Consequently, the checklist method cannot stand alone. A supplementing technique is needed in order to help answering the question of accuracy.

2.2 Comparative evaluation

Comparative evaluation is a method suited for answering the question of accuracy. If we try to assess the quality of something, we find that we often need something to compare with. How well does this wine taste? How beautiful is this scenery? How well does this dancer perform? These questions are difficult to answer without comparing with other wines, sceneries, or dancers. Comparative thinking plays a key role in almost all evaluation – also when it comes to evaluating sources on the Internet. Comparative evaluation is all about equating and verifying information from different sources. To illustrate the method, we will again borrow an example from Marc Meola [8] who has advocated the use of comparative evaluation over the checklist method.

Meola wanted to find out how many people got killed in the *Mai Lai massacre* during the Vietnam War. He searched for "mai lai" on Google and found a page reporting the body count to be 300. But can this number be trusted? To find out, Meola decided to double check with other sources. He found another page reporting the count to be 347; a third reported the count to be 500; a fourth reported the count to be 504. In other words, none of these numbers can be taken for granted. We have to consult authoritative sources if we want to know the real number of people who lost their lives. The point is that we only discover this if we double check the information we find. Comparison and verification are therefore essential elements when trying to answer the question of accuracy.

It will often be quite easy to double check information. The politician from New Zealand could easily have found out that dihydrogen monoxide is nothing but plain water by a few searches on Google. It would have taken just a little investigative reporting to find out that the Cypriote fans in question wear traditional sports clothes and sing traditional football songs. And when it comes to medical issues, it is of course the medical authorities that need to be consulted.

3. DOCUMENT TYPOLOGY AND AUTHORITY

Respect my authoritah!
Eric Cartman

From the previous examples, it is clear that some kind of authority hierarchy exists. Some sources seem to possess higher authority compared to others, and sources with higher authority seem to be more trustworthy than sources with lower authority.

In the field of History, researchers distinguish between first-hand accounts and second-hand accounts [9]. An account created by a person having experienced the reported incidences him- or herself is called a first-hand account. A second-hand account is reversely created by a person that has not experienced the reported incidences him- or herself. Instead, this person has gotten the reported information from other persons or media. All other things being equal, a first-hand account is viewed as more trustworthy than a second-hand account.

A similar hierarchy is found in scientific and scholarly communication. Information scientists have studied this social system of communication for many years, and found it to consist of a diversity of knowledge producers, intermediaries, institutions and users. Communication is taking place in both formal and informal channels, and is stored in different types of documents (e.g., articles, monographs, book reviews, conference proceedings, bibliographies, dictionaries, handbooks, encyclopedias and review articles to name a few). Information scientists have developed a number of models [e.g. 10, 11, 12, 13, 14] to describe this system, and from this derived a typology of documents and their functions. Typically, types of documents are grouped in three main categories:

- Primary literature
- Secondary literature
- Tertiary literature

Original research results published for the first time define the content of the first category (primary literature). Typical examples include journal articles, conference papers, doctoral dissertations and monographs. Critical book reviews and patents are usually also classified as primary literature.

Secondary literature types are derived from the primary literature. Textbooks, popular science monographs, handbooks, encyclopedias, dictionaries and review articles are among the most common types. Their function is to summarize, simplify, and communicate the results and knowledge produced in the primary literature.

Tertiary literature types include bibliographies, indexes, abstract services and catalogues. As such, they are aids to navigate the primary and secondary literature.

Just as the fist-hand account all other things being equal is more reliable than the second-hand account, original research results published in the primary literature are usually seen as more dependable than simplified summarizations communicated in the secondary literature. Consequently, when an information seeker needs to evaluate the trustworthiness of retrieved information, it is advisable to move upwards in the authority hierarchy. This is of course also

the case when it comes to information retrieved on the Internet. In the next section we shall see an example of how to climb *Mount Dependable* using all three categories of literature.

4. CLIMBING MOUNT DEPENDABLE

I like being near the top of a mountain.
One can't get lost here.
Wislawa Szymborska

As an active researcher I have a natural interest in scholarly publishing. I write and review papers for a number of academic journals and scientific conferences each year, and follow the new developments in my areas of expertise quite intensively. A few years ago I heard about a new 'thing' called *open access*. I learned that open access (OA) is the practice of providing unrestricted access via the Internet to scholarly publications. It struck me as an important development in scholarly publishing and I wanted to learn more about it, and about the possible consequences (positive or negative) of this new trend. Consequently, I accessed the Internet and made a search for the terms 'open access' on Google. Google is of course a search engine, and as such an index to information on the Internet. Technically, Google can therefore be categorized as a piece of tertiary literature.

One of the first links in the search result was a link to a page in the Danish version of Wikipedia [14]. Here I could read about the history of OA and about the various types of OA. Moreover, I was informed about a positive effect of OA, namely that authors of OA publications can expect to get cited much more than authors of non-OA publications:

> "Different studies have shown that OA papers typically get twice as many citations compared to papers that are not freely available" [my translation].

This OA-advantage postulate was quite a surprise for me. One of my biggest research interests is citation theory – an area of bibliometrics concerned with understanding what make authors of academic texts cite the sources they do. Historically, there have been two competing views on this known respectably as *the Mertonian view* and *the social constructivist view*.

Proponents of the Mertonian camp believe that science is a normative institution governed by internal rewards and sanctions. Scientists are believed to exchange information (in the form of publications) for recognition (in the form of awards and citations). This view suggests that citations are a way to acknowledge intellectual debts, and thus are mostly influenced by the worth as well as the cognitive, methodological, or topical content of the cited articles.

Instead of scientists using citations to impart recognition and to protect the property rights of a worthy piece of scholarship, social constructivists have portrayed scientists as using citations for their own convenience as tools of persuasion.

I conducted and published a detailed analysis of these two conflicting citation theories in my PhD-dissertation in 2004 [15; see also 16] and concluded that both of them lack empirical support. In short: Scientists are not as well-behaved as the Mertonian view prescribes them to be, nor are they as indolent and egoistic as portrayed by the social constructivist view. Thus,

learning now that OA papers get twice as many citations compared to non-OA papers made be doubt my own conclusions. If true - that is if OA-papers are really cited significantly more than non-OA papers just because they are freely available - then it clearly contradicts the Mertonian view and supports the social constructivist view. I was therefore eager to find out if it were really true or not.

Wikipedia describes itself as being a collaboratively edited, multilingual, free Internet encyclopedia that contains more than 25 million articles in 285 languages [17]. The articles are written collaboratively by volunteers worldwide, and almost all the articles can be edited by anyone with access to the site. As such, Wikipedia categorizes as secondary literature. Its main function is not to produce new knowledge or to validate existing knowledge claims, but instead to summarize, simplify, and communicate information from all branches of knowledge. The open nature of Wikipedia has led to various concerns regarding the accuracy of information. Wikipedia has responded to these concerns:

> "Some articles contain unverified or inconsistent information, […] though a 2005 investigation in Nature showed that the science articles they compared came close to the level of accuracy of Encyclopædia Britannica and had a similar rate of "serious errors" […]. Britannica replied that the study's methodology and conclusions were flawed, […] but Nature reacted to this refutation with both a formal response and a point-by-point rebuttal of Britannica's main objections […]" [17]

Yet, Wikipedia acknowledge that some articles contain errors, and I therefore decided to double check with other sources. Back in the Google results I found a link to an article from the peer reviewed OA journal PLoS Biology: *Citation advantage of open access articles* by Gunther Eysenabach [18]. The article reports the results of a longitudinal bibliometric analysis of a cohort of OA and non-OA articles:

> "Articles published as an immediate OA article on the journal site have higher impact than self-archived or otherwise openly accessible [...] articles. We found strong evidence that, even in a journal that is widely available in research libraries, OA articles are more immediately recognized and cited by peers than non-OA articles published in the same journal. OA is likely to benefit science by accelerating dissemination and uptake of research findings" [18].

This seemed to support the OA-advantage postulate. Having now climbed to the highest level of the authority hierarchy (i.e. the primary literature) I was close to reject my former conclusions. Yet, a few things made me hesitate. First, who is this Eysenbach? Is he really qualified as a bibliometrician? Secondly, could it be that an OA-journal like PLoS Biology has an obvious interest in supporting the OA-advantage postulate? Could this perhaps influence on the peer review process and perhaps make reviewers less critical? Third, does this study really support the OA-advantage postulate? Is it the freely available status alone that caused the reported citation effect or could it be that other factors are the main contributors?

My first two concerns were quickly dismissed. Gunther Eysenbach is an MD and works as a senior scientist at the Department of Health Policy, Management and Evaluation, University of Toronto [19]. His publication list includes publications in many respected journals (e.g. Lancet; British Medical Journal; JAMA), and he has written a number of bibliometric articles.

PLoS Biology is a highly respected scientific journal with a 2011 impact factor of 11.452, ranking it first in the Biology category [20].

To address my third concern I decided to check if someone had commented on Eysenbach's results. I consequently turned to Google Scholar and made a search for articles having cited Eysenbach's results. One of the first results linked to an article from Journal of Informetrics (JoI): *Do open access articles have greater citation impact?* The article is written by five authors representing Wiley-Blackwell, Elsevier and Thomson Scientific [21]. It presents a critical review of the literature examining the relationship between OA status and citation counts of scholarly articles. The authors find that:

> "Three non-exclusive postulates have been proposed to account for the observed citation differences between OA and non-OA articles: an open access postulate, a selection bias postulate, and an early view postulate" [21].

After having critically examined the available literature they conclude:

> "The most rigorous study to date (in condensed matter physics) showed that, after controlling for the early view postulate, the remaining difference in citation counts between OA and non-OA articles is explained by the selection bias postulate. No evidence was found to support the OA postulate per se; i.e. article OA status alone has little or no effect on citations" [21].

Thus, having now climbed to the very top of the authority hierarchy I concluded that the Wikipedia article was wrong about OA articles being cited twice as much as non-OA articles.

When writing this paper it suddenly struck me that perhaps the five authors of the JoI-article was somehow influenced by their institutional affiliation. After all, they represent academic publishers that could have an interest in arguing against the OA movement. Yet, there is a fine line between vigilance and paranoia which one should not cross. Moreover, JoI is an international peer reviewed journal widely respected as a dependable source for bibliometric information. I therefore choose to keep believing that Wikipedia got it wrong and JoI got it right.

4. CONCLUSION

Change is the end result of all true learning.
Leo Buscaglia

The Internet is truly fantastic! However, the information we find on the Internet must be used with care. We can to some extent check the validity of information using the so-called checklists, but they cannot stand alone. When it comes to important information we should always seek to double check and verify the validity by comparing what other sources say. In this process it is important to know the so-called authority hierarchy. Not all sources possess the same level of authority. Distinguishing between three levels of authority (tertiary literature, secondary literature, and primary literature) it is recommended that information seekers strive to double check and verify information on a higher authority level – preferably the highest (the primary literature). This is of course not always possible. Reading the primary

literature on highly specialized topics often requires special skills that only the specialist masters. Yet, one should always seek to verify important information to the highest level one can understand, and in special cases be willing to seek advice from experts.

5. LINKS* AND REFERENCES

*All links are checked March 14. (2013).

[1] http://www.stuff.co.nz/national/politics/38005/National-MP-falls-victim-to-water-hoax

[2] http://www.b3ta.com/links/Lazy_Journalist

[3] http://whatstheharm.net/internetmisinformation.html

[4] Popper, K. (1995). *The Logic of Scientific Discovery*. London, UK: Routledge.

[5] Merton, R.K. ([1942] 1973). The normative structure of science. In: Merton, R.K. (ed.), *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago, IL: University of Chicago Press: 267-278.

[6] A brief selection of checklists:
http://lib.nmsu.edu/instruction/evalcrit.html
http://aics.education.wisc.edu/LessonPlans/FirstAmericansGreatLakesRegion/credcheck.pdf
http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/EvalWorksheet.pdf
http://infopeople.org/resources/internet/evaluatingchecklist
http://www.andyspinks.com/researchhelp/web/CARS.pdf
http://www.montgomeryschoolsmd.org/schools/wjhs/depts/socialst/Media/internet.html

[7] http://www.martinlutherking.org/articles/BeastAsSaint.pdf

[8] Meola, M. (2004). Chucking the checklist: A contextual approach to teaching undergraduates web-site evaluation. *Portal – Libraries and the Academy*, 4(3): 331-344

[9] Howell, M. & Prevenier, W. (2001). *From Reliable Sources: An Introduction to Historical Methods*. Ithaca, NY: Cornell University Press.

[10] Crawford, S.Y., Hurd, J.M. & Weller, A.C. (1996). *From Print to Electronic: The Transformation of Scientific Communication*. Medford, NJ: Information Today.

[11] Garvey, W.D. & Griffith, B.C. (1972). Communication and information processing within scientific disciplines :Empirical findings for psychology. *Information Storage & Retrieval*, 8: 123-136.

[12] Fjordback Søndergaard, T., Andersen, J. & Hjørland, B. (2003). Documents and the communication of scientific and scholarly information: Revising and updating the UNISIST model. *Journal of Documentation*, 59(3): 278-320

[13] Subramanyam, K. (1981). *Scientific and Technical Information Resources*. New York, NY: Marcel Dekker.

[14] http://da.wikipedia.org/w/index.php?title=Open_access&oldid=1163222

[15] Nicolaisen, J. (2004). *Social Behavior and Scientific Practice – Missing Pieces of the Citation Puzzle*. Copenhagen, DK: Royal School of Library and Information Science.

[16] Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology*, 41: 609-641.

[17] http://en.wikipedia.org/wiki/Wikipedia

[18] Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biology*, 4(5): http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0040157

[19] http://www.hctp.utoronto.ca/PeopleMentorsDetails.asp?pRID=106

[20] *Journal Citation Reports*: http://wokinfo.com/products_tools/analytical/jcr/

[21] Craig, I.D., Plume, A.M., McVeigh, M.E., Pringle, J. & Amin, M. (2006). Do open access articles have greater citation impact?: A critical review of the literature. *Journal of Informetrics*, 1(3): 239-48.